CROMER: a Tool for Cross-Document Event and Entity Coreference

Christian Girardi[1], **Manuela Speranza**[1], **Rachele Sprugnoli**[1-2], **Sara Tonelli**[1]
[1]Fondazione Bruno Kessler, Via Sommarive 18, 38123 Povo (TN), Italy
[2]University of Trento, Via Sommarive 5, 38123 Povo (TN), Italy
E-mail{cgirardi | manspera | sprugnoli | satonelli}@fbk.eu

Abstract

In this paper we present CROMER (*CROss-document Main Events and entities Recognition*), a novel tool to manually annotate event and entity coreference across clusters of documents. The tool has been developed so as to handle large collections of documents, perform collaborative annotation (several annotators can work on the same clusters), and enable the linking of the annotated data to external knowledge sources. Given the availability of semantic information encoded in Semantic Web resources, this tool is designed to support annotators in linking entities and events to DBPedia and Wikipedia, so as to facilitate the automatic retrieval of additional semantic information. In this way, event modelling and chaining is made easy, while guaranteeing the highest interconnection with external resources. For example, the tool can be easily linked to event models such as the Simple Event Model [Van Hage et al, 2011] and the Grounded Annotation Framework [Fokkens et al. 2013].

Keywords: cross-document coreference, annotation tool, Semantic Web

1. Introduction

Developing a tool for cross-document event and entity coreference is challenging for several reasons, both from the technical and the conceptual point of view. The main issue is the fact that no consolidated standard for event coreference annotation has been established in the NLP community.

The MUC approach to coreference has been criticized for mixing anaphora with other coreference phenomena [van Deemter and Kibble, 1995]. The same conflation is observed in the ACE (*Automatic Content Extraction*) program datasets, in which annotators perform intra-document coreference by grouping all mentions of the same entity, be it named, nominal or pronominal mentions (see the latest version of the guidelines [Linguistic Data Consortium, 2008a]). In 2008, a cross-document global integration and reconciliation of information on annotation has also been performed within the ACE evaluation initiative, but only for 50 person and organization entities and only for documents in which the target entities of interest were mentioned by name [Linguistic Data Consortium, 2008b]. As for event coreference, in ACE 2004 evaluation the event detection and linking task was included for the first time but only at the intra-document level [Linguistic Data Consortium, 2004b]. Within the recent OntoNotes annotation, noun phrases, nominals (but not adjectival pre-modifiers) and verbs can be marked as co-referent [BBN Technologies, 2011] but only in an intra-document perspective. In particular, two types of coreference chains are marked, namely appositive constructions (e.g. *the PhacoFlex intraocular lens, the first foldable silicone lens available for cataract surgery*) and anaphoric coreference (e.g. *Elco Industries Inc. said it expects net income in the year ending June 30, 1990*).

More recently, researchers started to develop resources in which events are annotated across multiple documents, such as the EventCorefBank [Bejan and Harabagiu, 2010]. Cross-document coreference is challenging also because it is not straightforward to identify the trigger event in the chain of events. Descriptions of events across documents may complement each other providing a complete picture, but still textual descriptions tend to be incomplete and sparse with respect to time, place and participants. At the same time, the comparison of events becomes more complex. [Nothman et al., 2012] proposes to relax the notion of coreference taking into consideration only the linking between an event reference and the target news story where the event was reported for the first time. Although they still report a low inter-annotator agreement on which tokens are to be linked (minor than 0.30), the agreement on the link target for agreed tokens shows to be substantial (0.73).

With CROMER, the problem of finding the trigger event is tackled in a completely different way: we rely on an external semantic representation of the event, which we call *event instance*, and we link each *mention* (intra- and cross-document) to it. This instance is possibly linked to DBPedia or any other knowledge base used by the annotators and is uniquely identified by time, place and participants. For each of such instances, a template is created in the CROMER tool for top-down event coreference. The same approach has been adopted with entities, distinguishing between entity mentions in text and their formal representation as entity instances in a semantic layer.

As far as manual annotation tools for cross-document annotation are concerned, to our knowledge the literature reports only about the EDNA plugin for Callisto [Day et al., 2008] and the web interface designed for cross-document coreference resolution of Italian person entities within the OntoText project [Bentivogli et al., 2008]. Similarly to CROMER, Callisto/EDNA is based on a Tomcat web server and on a Lucene document parser. On the other hand, Callisto/EDNA has three constraints

that differentiate it from CROMER: first of all its use is strongly dependent on previously annotated corpora following the intra-document ACE Entity Detection and Recognition guidelines. Second, it does not allow cross-document annotation of events. Moreover the enrichment of the annotation with semantic linking to an external knowledge base is not provided. As for the OntoText interface, it allows multi-user web annotation like CROMER but it has been developed to process only person named entities and no linking of the annotated data to Semantic Web resources is possible.

To summarize, the advantages offered by CROMER are manifold. First, it is a tool that can deal both with events and entities, overcoming the need to have different annotation systems for the two elements. Then, it has been designed following a top-down approach, namely starting from the definition of a *template* describing the event or entity instance, and then linking it to mentions in text. This overcomes the issue of choosing a trigger event in the document that starts the coreference chain. Then, the fact that it is based on templates makes it easy to integrate the annotated data with semantic web resources, thanks also to the possibility to connect each template with an item from an external knowledge base (typically DBpedia). Lists of templates can also be imported by the user, taking advantage of the availability of structured data. However, with CROMER it is also possible to import documents annotated with intra-document coreference, thanks to the full compatibility with the Content Annotation Tool (CAT) [Bartalesi Lenzi et al., 2012]. In this case, the data exported from CAT can be directly imported in CROMER and the user is required to add the inter-document coreference layer.
The flexibility of the tool and the fact that is satisfies the needs both of linguists and of semantic web experts is an outcome of the complex process that has led to the development of the tool, in which researchers from different groups have been involved and invited to provide feedback on the tool functionalities.
More details on the single items mentioned above will be provided in the following sections.

2. Annotation Workflow

Annotation with CROMER has the aim of marking-up coreference between entities and between events across different documents. While intra-document coreference is a well-established field of research at least for entities, the work on cross-document coreference is still burgeoning especially for events [Bejan and Harabagiu, 2010] [Lee et al., 2012]. Our approach is to combine textual information with information taken from external knowledge sources (such as DBpedia) through a manual linking performed by annotators. The use of external sources of information makes it possible to correctly establish the fact that two or more expressions refer to the same entity or to the same event. The goal of the annotation is to associate to an event or an entity a set of documents, where such event/entity is mentioned at least once. Annotation at mention level (intra-document) is also supported but not mandatory.
CROMER allows for collaborative annotation, as different annotators (logged in as individual users) can use the tool to work on the same set of documents; in this case, all instances and templates are shared among them. Two annotation modes are available: in the first case, annotators cannot see each other's annotations (which is useful for annotating data to be used for inter-annotator agreement computation), while in the second mode the annotated data are shared and can be seen by all annotators. When needed, a "judge" annotator can solve discrepancies and modify existing annotations by logging in as administrator.
The annotation workflow is top-down and comprises the following steps:

1. A seed set *S* of entities or events of interest is defined by the annotator.
2. For each entity and event instance in *S*, a generic template has to be filled in (see Figure 1). This is done once before starting the annotation.

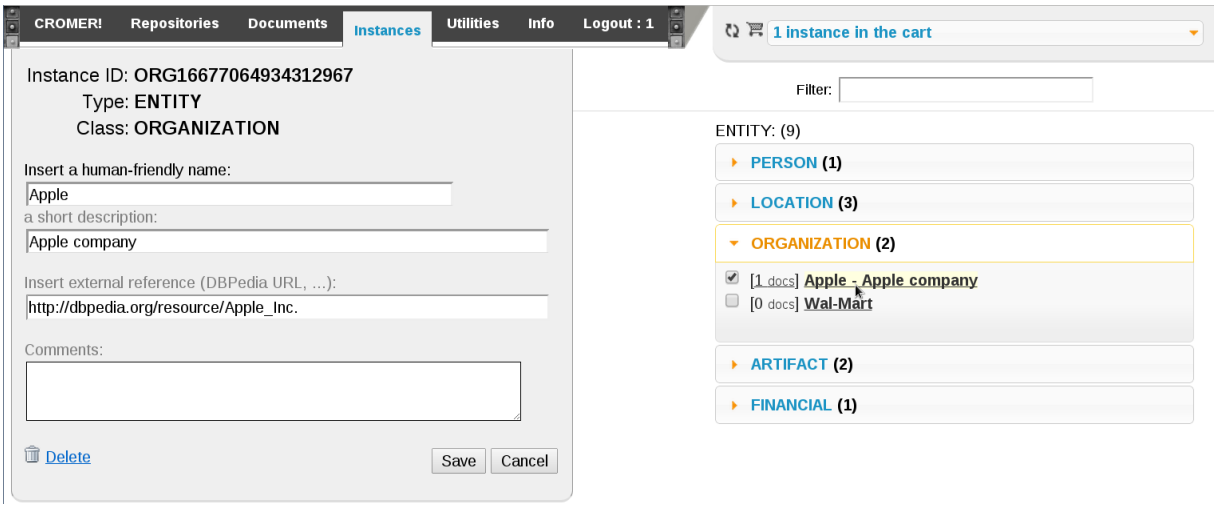

Figure 1: The template of an entity

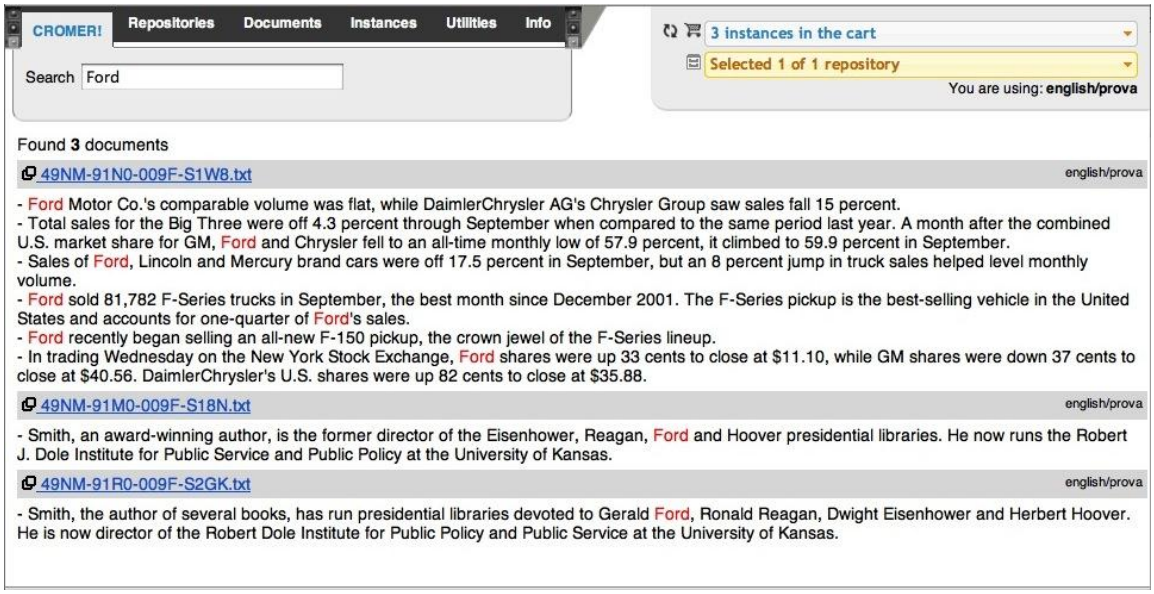

Figure 2: Documents returned after performing a search by word

3. The annotator uploads a collection of documents that she/he considers relevant to the topic she/he wants to annotate, and that may contain the elements in *S*.
4. The annotator searches the collection of documents for the entities in *S*. A string-based search is possible. Some thresholds can be set, for instance (i) the seed elements must occur at least in *n* different documents of the cluster, (ii) and the clusters cannot include more than *n* documents. The first constraint has been adopted to obtain interesting instances for cross-document coreference, whereas the second restriction helps avoiding that annotation is too time consuming. The search returns a subsection D of the document collection (see Figure 2).
5. The annotator checks each *d* in *D* to see if the event/entity instance mentioned in each document corresponds to the one described in the template. If not, the documents is discarded. In this phase, it is not necessary to check all mentions in the document, one is considered enough to include *d* in the final entity/event cluster.
6. After all documents in *D* have been validated, it is possible to export the final entity/event cluster, containing only documents in which the event/entity in *S* is mentioned.

The tool has no external dependencies and intra-document annotation is not required. However, CROMER is compatible with the Content Annotation Tool (CAT) [Bartalesi Lenzi et al., 2012] for input format. This allows to perform cross-document annotation on top of the intra-document annotation performed with CAT, thus merging top-down and bottom-up information into a single representation.

More specifically, it is possible to import from CAT:

- automatic sentence splitting and tokenization;
- manual annotation of events, mentions, and co-reference relations at the intra-document level[1].

If intra-document co-reference chains are imported from CAT, in Step 5 the annotator has the possibility to visualize all the mentions of a CROMER instance occurring in a document and to assign them all to that instance by simply acting on the co-reference chain (see Figure 3).

Templates for entity and event instances contain different types of information. Since Wikipedia and DBPedia are based on concepts, which are typically expressed by nouns, nominal entities are usually found in such resources and can be easily linked to a template. In the case of verbal events, on the other hand, it is more difficult to find that specific event instance, rather than a generic notion. For instance, the template of the 'tsunami striking Indonesia in 2004' event instance should not be linked to the DBPedia page on 'tsunamis'[2], but to the page describing this event having a precise location in time and space[3].

Fields related to *entities* are the following:

- id, a number that uniquely identifies the entity, automatically generated by the annotation tool;
- name, a human-friendly identifier of the entity;
- link, URI taken from an external knowledge base (e.g. DBpedia);
- class, corresponding to different semantic classes, e.g. person and location.

[1] Through the CROMER configuration file it is possible to customize the list of markables and co-reference relations to be imported from CAT.

[2] http://dbpedia.org/page/Tsunami

[3] http://dbpedia.org/page/2004_Indian_Ocean_earthquake_and_tsunami

Figure 3: Assignment of mentions annotated in CAT to a CROMER instance

The following fields are assigned to each *event*:

- id, number that uniquely identifies the event, automatically generated by the annotation tool;
- name, a human-friendly identifier of the event;
- link, URI taken from an external knowledge base (e.g. DBpedia);
- class, corresponding to different semantic classes of events, such as *communication* and *cognitive* events.

Some strategies are implemented to speed up manual annotation and take in input pre-processed data. These strategies include:

1. The way document clusters are produced: in order to speed up annotation, it is possible to feed the system with a document collection (a folder in gzipped format) and an external file, where several file names are associated with different entity/event clusters. For instance, it is possible to define in this file which documents in the uploaded folder mention "Volkswagen A.G.", and which ones mention "Porsche". Overlaps among different clusters are also allowed.
2. The way templates are defined: templates can be created within the application, but also imported in a specific format, so as to reduce manual effort to retrieve information on events/entities and linking to external sources. In the future, also the possibility to establish relations between templates will be implemented.

3. Implementation details

CROMER has been developed in Java as a TomCat web application. The data are stored in several Lucene indexes (one for each document repository and one for all user-defined instances of events and entities) and a MySQL database (for the user annotations). In order to avoid consistency problems during the export of the data, a check to compare the Lucene indexes and the content of the database is performed.

Utilities implemented in CROMER include:

- *Import* functionalities:

A user can *import instances* and corresponding templates from a file in tab separated format. This file should include fields such as instance ID, type, class, naming, etc. If an imported instance is already present in the database, it will be recognized based on the instance ID and updated, otherwise it will be created from scratch.

Another functionality is the import of *documents*, which can be saved by the user in separated repositories. Supported formats are raw text, tokenized text and CAT XML files. In case of raw text documents, automatic built-in tokenization for English and Italian is performed by TokenPro [Pianta et al., 2008]. Tokenizers for different languages can be added as well.

- *Search* and *retrieval* functionalities:

CROMER supports *document-based* search using single words, strings of tokens and wildcards (see Fig. 2). The search can cover all imported documents or target specific repositories. The search can also start from an instance and display all documents already annotated with mentions of such instance.

Another search type is *instance-based*, i.e. a user can retrieve all instances matching a specific string inside the instance repository. Through this search a user can select a set of already existing instances to be annotated in new documents.

Thanks to the Lucene indexes, the search functionalities described above are very efficient also with large repositories of documents and instances.

- *Export* functionalities:

CROMER supports the export of documents in tab separated format and in CAT format. This latter feature enables users to start annotating intra-document coreference with CROMER and then work at document level with CAT. A user can also export the list of instances in csv format, manually change or enrich it and then import again the list into the tool.

Other technical features include:

- Automatic validation of external reference links when creating or modifying instance templates. For DBpedia URIs, an additional autocomplete control has been added.
- Statistics on the performed annotations specific to

single users (number of annotated documents, instances and associated mentions);

- User profiling: users can have different permission according to their role (admin or other). An administrator can configure some system preferences (e.g. color management and import settings), and create new user accounts.

CROMER is released under Apache license and it is distributed on GitHub at the following URL: http://github.com/hltfbk/CROMER/. It is a free open-source software, which can be downloaded, installed locally and easily customized by the user.

We made some preliminary analyses on the activity of two expert annotators to track their speed when using the tool [Cybulska and Vossen, 2014]. We observed that the average time needed to perform the annotation of a mention inside the document (i.e. select the mention and connect it to the entity instance) is around 20 seconds, averaged over 4,000 assignments. We did not record any particular issue or anomaly during the annotation workflow.

4. Conclusion

In this paper we presented CROMER, a tool for cross-document coreference. To our knowledge, this is the only tool available (as open-source software) for both event and entity annotation. CROMER offers several functionalities, such as the possibility to annotate in a top-down fashion starting from event and entity templates, and its full compliance with the Content Annotation Tool (CAT) for intra-document coreference. Besides, it has been designed so as to satisfy the requirements of the Semantic Web community by integrating the possibility to link the templates to external knowledge sources (e.g. DBpedia).

Our plans for future work include several aspects. In particular, we will implement the possibility to create specific relations between two entities (e.g. EntityA *member_of* EntityB), two events (e.g. EventA *sub_event_of* EventB), or between an event and an entity (e.g. EventA *has_participant* Entity B). We will also enable users to create new template fields and modify existing ones through the CROMER interface. Finally, we will give users the possibility to export documents in NAF (NLP Annotation Format), an XML-based format designed to represent linguistic annotations in complex NLP pipelines [Fokkens et al., 2014].

5. Acknowledgements

This research is supported by the European Union's 7th Framework Programme via the NewsReader Project (ICT-316404).

6. References

Linguistic Data Consortium. (2004). The ACE 2004 Evaluation Plan. Technical report, LDC.

Linguistic Data Consortium (2008a). ACE (Automatic Content Extraction) English Annotation Guidelines for Entities, Version 6.6 2008.06.13.

Linguistic Data Consortium. (2008b). ACE 2008: Cross-Document Annotation Guidelines (XDOC), Version 1.6 -- 2008.05.06.

BBN Technologies (2011). Co-reference Guidelines for English OntoNotes Version 6.0, http://catalog.ldc.upenn.edu/docs/LDC2007T21/coreference/english-coref.pdf

Bartalesi Lenzi, V, Moretti, G, Sprugnoli, R, 2012 CAT: the CELCT Annotation Tool. In *Proceedings of LREC 2012*, Istanbul, Turkey.

Bejan, C. A., Harabagiu, S. M. (2010). Unsupervised Event Coreference Resolution with Rich Linguistic Features. In Jan Hajic; Sandra Carberry & Stephen Clark, eds., *Proceedings of ACL'10* (pp. 1412--1422), The Association for Computer Linguistics.

Bentivogli, L., Girardi, C., Pianta, E. (2008). Creating a Gold Standard for Person Cross-Document Coreference Resolution in Italian News. In *Proceedings of the LREC 2008 Workshop on Resources and Evaluation for Identity Matching, Entity Resolution and Entity Management*. Marrakech, Morocco, May 31st 2008.

Cybulska, A., Vossen, P. (2014) Using a sledgehammer to crack a nut? Lexical diversity and event coreference resolution. In *Proceedings of The 9th edition of the Language Resources and Evaluation Conference (LREC)* , 26-31 May, Reykjavik, Iceland.

Day, D., Hitzeman, J., Wick, M. L., Crouch, K., Poesio, M. (2008). A Corpus for Cross-Document Co-reference. In *Proceedings of the LREC 2008*. European Language Resources Association (ELRA).

Fokkens, A, van Erp, M., Vossen, P., Tonelli, S., van Hage, W., Serafini, L., Sprugnoli, R., Hoeksema, J. (2013). GAF: A Grounded Annotation Framework for Events. In *Proceedings of 1st EVENTS Workshop*, Atlanta, US.

Fokkens, A., Soroa, A., Beloki, Z., Rigau, G., van Hage, W.R., and Vossen, P. (2014) NAF: the NLP Annotation Format. NWR-2014-3. VU University Amsterdam.
http://www.newsreader-project.eu/files/2013/01/techreport.pdf

Lee, H., Recasens, M., Chang, A. X., Surdeanu, M., Jurafsky, D. (2012). Joint Entity and Event Coreference Resolution across Documents. In *Proceedings of EMNLP-CoNLL* (pp. 489--500). The Association for Computer Linguistics.

Nothman, J., Honnibal, M., Hachey, B., Curran, J. R. (2012). Event Linking: Grounding Event Reference in a News Archive. In *Proceedings of ACL'12* (pp. 228--232). The Association for Computer Linguistics.

Pianta, E., Girardi, C., Zanoli, R. (2008). The TextPro Tool Suite. In *Proceedings of the LREC 2008*. European Language Resources Association (ELRA).

van Deemter, K., Kibble, R. (2000). On Coreferring: Coreference in MUC and Related Annotation Schemes. *Computational Linguistics* 26 (4) , 629--637.

van Hage, W. R., Malaisé, V., Segers, R., Hollink, L. (2011). Design and Use of the Simple Event Model. *In Journal of Web Semantics*, vol.9, n. 2